# Determination of Protein Secondary Structure Using Factor Analysis of Infrared Spectra[†]

David C. Lee,*[‡] Parvez I. Haris,[§] Dennis Chapman,[§] and Robert C. Mitchell[‡]

*SmithKline Beecham Pharmaceuticals, The Frythe, Welwyn, Hertfordshire AL6 9AR, U.K., and Department of Protein and Molecular Biology, Division of Basic Medical Science, Royal Free Hospital School of Medicine, Rowland Hill Street, London NW3 2PF, U.K.*

*Received October 5, 1989; Revised Manuscript Received June 1, 1990*

ABSTRACT: A method is presented for determining the secondary structural composition of a protein in aqueous solution from its infrared spectrum. A factor analysis approach is used to analyze the infrared spectra of 18 proteins whose crystal structures are known from X-ray studies. Factor analysis followed by multiple linear regression identifies those eigenspectra that correlate with the variation in properties described by the calibration set. The properties of interest in this study are % $\alpha$-helix, % $\beta$-sheet, and % turns. In the analysis of an unknown, the factor loadings required to reproduce its spectrum are substituted in the regression equation for each property to predict its secondary structural composition. The accuracy of the method was determined by removing each standard, in turn, from the calibration set and using a calibration set generated from the remainder to predict its composition. By this method we obtain standard errors of prediction of 3.9% for $\alpha$-helix, 8.3% for $\beta$-sheet, and 6.6% for turns. The method may also be applied to the spectra of proteins in $^2H_2O$. The method has important advantages over those currently in use for the quantitative analysis of the infrared spectra of proteins. Manipulation of the spectrum is kept to a minimum, no curve-fitting is necessary, and the several amide I band components need not be assigned.

Infrared spectroscopy is now a widely used technique for the analysis of protein secondary structure. With the advent of Fourier transform methods for data acquisition came the ability to examine dilute aqueous solutions via signal averaging and digital subtraction of background absorptions. Common features of the infrared spectra of proteins are the so-called "amide" bands, which arise from delocalized vibrations of the peptide linkage (Susi, 1969). Of these, the amide I band (1700–1600 cm$^{-1}$) is the most useful for the analysis of protein secondary structure (Susi et al., 1967; Timasheff et al., 1967; Susi, 1969). Correlations of amide I band frequency with the presence of $\alpha$-helical, antiparallel and parallel $\beta$-sheet, and random coil structures in proteins in $^2H_2O$ (Susi, 1969) and $H_2O$ (Koenig & Tabb, 1980) are well established. The widths of the amide I bands assigned to the various secondary structures are large compared to the separation of their peak maxima, so that the amide I absorption of a complex protein structure consists of several overlapping bands. The application of algorithms for the visualization of overlapping bands, such as Fourier self-deconvolution (Kauppinen et al., 1981) and second derivatives (Maddams & Southon, 1982), have enabled the identification and assignment of $\alpha$-helical, $\beta$-sheet, and turn structures in soluble proteins (Susi & Byler, 1983) and membrane proteins (Lee & Chapman, 1986; Lee et al., 1987).

Derivative and deconvolution analysis of the infrared spectra of proteins provides information of a qualitative nature. Quantitative estimations of secondary structure may not be made directly from peak intensities or band areas in derivative spectra or peak intensities in deconvolved spectra, since these parameters are heavily influenced by differences in the original bandwidths of the overlapping components. Early attempts at quantitative analysis involved curve-fitting directly to the amide I band (Ruegg et al., 1975; Eckert et al., 1977). More recent studies have used derivatives and deconvolution to identify the positions of overlapping peaks followed by iterative curve-fitting to either the deconvolved spectra (Susi et al., 1985; Byler & Susi, 1986; Surewicz et al., 1987; Holloway & Mantsch, 1989) or the original difference spectra (Yang et al., 1987; Arrondo et al., 1988). The most detailed of these studies (Byler & Susi, 1986) compared the percentages of $\alpha$-helix and $\beta$-sheet for 17 proteins in $^2H_2O$ as determined by infrared spectroscopy with the X-ray crystallographic assessments of Levitt and Greer (1977). While the general agreement between the X-ray and infrared values was good, there are a number of difficulties with this approach [see Discussion and Mantsch et al. (1989)] if a consistent application between laboratories is to be achieved.

As an alternative, we decided to investigate the ability of factor analysis to provide quantitative data from the Fourier transform infrared (FTIR)[1] spectra of proteins. Factor analysis is widely employed in chemistry to provide both qualitative and quantitative information. A description of the principles, mathematics, and applications of factor analysis is provided in the monograph by Malinowski and Howery (1980). A major advantage of this approach is that large quantities of data of great complexity can be analyzed. When factor analysis is applied in spectroscopy, only minimal interpretation of the original data is necessary, and additionally, the complete spectrum can be analyzed in order to identify bands associated with particular components or properties. Programs for factor analysis are now supplied by many of the spectrometer manufacturers. As we shall show, this approach can be used to provide a novel, reliable, and reproducible means of extracting quantitative information from the FTIR spectra of proteins.

[†] We thank the Science and Engineering Research Council and the Wellcome Trust for financial support (P.I.H. and D.C.).

[‡] SmithKline Beecham Pharmaceuticals.

[§] Royal Free Hospital School of Medicine.

[1] Abbreviations: FTIR, Fourier transform infrared; SEP, standard error of prediction; *r*, product–moment correlation coefficient.

0006-2960/90/0429-9185$02.50/0

## Materials and Methods

Most protein samples were obtained from Sigma Chemical Co. Ltd. (Poole, U.K.) and were used without further purification. These were alcohol dehydrogenase (equine liver, A6128), calmodulin (bovine brain, P2277), carbonic anhydrase (bovine erythrocyte, C7500), $\alpha$-chymotrypsin (bovine pancreas, C7762), $\alpha$-chymotrypsinogen (bovine pancreas, C4879), concanavalin A (*Canavalia ensiformis*, C7275), cytochrome *c* (equine heart, C7752), elastase (porcine pancreas, E0258), hemoglobin (bovine erythrocyte, H2500), insulin (porcine pancreas, I3505), lysozyme (chicken egg white, L6876), myoglobin (sperm whale skeletal muscle, M0380), nuclease (*Staphylococcus aureus*, N3755), papain (*Papaya latex*, P4762), pepsin (porcine stomach, P6887), prealbumin (human plasma, P7528), protease (*Streptomyces griseus*, P0652), ribonuclease A (bovine pancreas, R5500), ribonuclease S (bovine pancreas, R6000), trypsin (porcine pancreas, T0134), and trypsinogen (bovine pancreas, T1143). Trypsin inhibitor (bovine pancreas) was a gift from Bayer AG. Porcine phospholipase $A_2$ was a gift from Dr. G. H. de Haas (State University of Utrecht).

For infrared spectroscopy, samples were prepared as 5% (w/v) solutions in water or 50 mM $KH_2PO_4$ buffer, pH 7.0 (ribonucleases A and S, lysozyme). For studies in deuterium oxide, 5% (w/v) solutions were prepared in $^2H_2O$ from Sigma Chemical Co. (99.8 atom % $^2H$, D4501). Hydrogen–deuterium exchange was followed by recording spectra after various periods of incubation at room temperature. Spectra were selected for analysis when hydrogen–deuterium exchange was judged to be complete or at equilibrium as determined by the absence of further spectroscopic changes in the region of the amide II band.

Infrared spectra were recorded by using a Perkin–Elmer 1750 FTIR spectrometer equipped with a fast-recovery TGS detector and a Perkin-Elmer data station. Aqueous samples were placed in a thermostated Specac 20500 cell fitted with either a 6-$\mu$m tin spacer (studies in $H_2O$) or a 50-$\mu$m Teflon spacer (studies in $^2H_2O$). The temperature of the sample was maintained at 20.5 ± 0.1 °C by means of a cell jacket of circulating water. The spectrometer was continuously purged with either dry air or dry nitrogen to eliminate water vapor absorption from the spectral region of interest. A sample shuttle was used to allow the background spectrum to be signal-averaged over the same time period as the sample spectrum. For samples in $H_2O$, 100, 200, or 256 scans were coadded, apodized with a medium Norton–Beer function, and Fourier transformed to give a resolution of 2 $cm^{-1}$. For samples in $^2H_2O$, 100 or 128 scans were coadded and processed as for the samples in $H_2O$.

Spectra of $H_2O$, $^2H_2O$, and buffer were recorded in the same cell and under the same instrument conditions as the sample spectra. Difference spectra were generated by an interactive difference routine to subtract the appropriate solvent spectrum from the spectrum of each protein solution. Proper subtraction of water was judged to yield an approximately flat baseline from 1900 to 1720 $cm^{-1}$, avoiding negative side lobes, and the removal of the water band near 2130 $cm^{-1}$. Subtraction of $^2H_2O$ was adjusted to the removal of the $^2H$–O–$^2H$ bending absorption near 1220 $cm^{-1}$. Second-derivative spectra were generated from the difference spectra by using the Perkin-Elmer DERIV routine. This utilizes the Savitzky–Golay derivative routine (Savitzky & Golay, 1964)—a 13-data-point (13-$cm^{-1}$) window was selected. Spectral deconvolution was performed by using Perkin-Elmer's ENHANCE function, which is analogous to the method of Kauppinen et al. (1981). Deconvolution parameters used for the amide I band were $\sigma$ = 6 $cm^{-1}$ (half-width at half-height) and $K$ = 2.0 (relative reduction in bandwidth).

*Quantitative Analysis.* Quantitative analysis of the secondary structure of a range of soluble proteins was made by using the program CIRCOM (Computerized Infrared Characterization of Materials), available from Perkin-Elmer Ltd. A full mathematical description of this method has been given (Fredericks et al., 1985b). Initially, a calibration set is generated from the IR spectra of a range of samples for which the properties of interest have been measured by other methods. In our case, the properties of interest are % $\alpha$-helix, % $\beta$-sheet, and % turn structures as determined by X-ray crystallography (Levitt & Greer, 1977). The IR spectra of 5% (w/v) solutions of 18 proteins were used to create the calibration set. CIRCOM utilizes factor analysis to generate abstract factors or eigenspectra, which may be combined linearly to enable the original spectra to be reconstructed within experimental error. Those factors that account only for noise in the original spectra are discarded. The remaining factors are analyzed for their contributions (factor loadings) to each of the spectra in the calibration set. Multiple linear regression is then used to establish correlations between these factor loadings and the composition of the calibration samples in terms of the properties of interest. For each property, those factors that show little correlation are eliminated and the regression is repeated until all remaining factors show significant correlation. The regression equation for each property is used to determine the contribution of that property to each of the spectra in the calibration set, and these values are compared with the "true" values (in this case, the X-ray values). In the analysis of the spectrum of an unknown, the factor loadings required to reproduce the spectrum are determined by using the factors generated from the calibration set. The value for each property in the unknown is then obtained by substituting in the relevant regression equation.

The importance of various spectral parameters in constructing the calibration set was investigated. These were, for difference spectra, the spectral range, normalization of the area under the amide I band (1700–1600 $cm^{-1}$), and normalization of the ordinate value at 1700 $cm^{-1}$. Additionally, calibration sets were generated by using both second-derivative spectra and deconvolved spectra calculated from the normalized spectra. The suitability of these various methods was evaluated via the product–moment correlation coefficients ($r$) generated by linear regression of the predicted and true values for each property. In order to assess the significance of a difference in correlation coefficient, the following procedure was applied. In those cases where the number of factors (i.e., terms in the regression equation) used to predict a type of structure was the same between calibration sets, the set with the smallest deviation sum of squares was chosen as the best model. The deviation sum of squares was calculated as the sum of the squares of the differences between CIRCOM and X-ray values for the calibration set. In those cases where CIRCOM achieves an increase in the correlation coefficient (and a decrease in the deviation sum of squares) by including more factors (terms) in the regression equation, an F-test was performed to evaluate if the more complex equation explained a significantly larger amount of the variation in predictions.

## Results

### (A) Spectra in $H_2O$

*Generation of Calibration Sets.* Difference spectra of 18 proteins, generated by the interactive digital subtraction of

Table I: Effect of Spectral Normalization on the Correlation Coefficients ($r$) for Helix, Sheet, and Turn Structures for the 18 Proteins[a]

| | $r$(helix) | $r$(sheet) | $r$(turns) | no. of factors[b] |
|---|---|---|---|---|
| A | 0.946 | 0.880 | 0.944 | 9 |
| B | 0.907 | 0.873 | 0.924 | 10 |
| C | 0.986 | 0.952 | 0.848 | 9 |
| D | 0.995 | 0.989 | 0.903 | 11 |
| E | 0.992 | 0.968 | 0.929 | 9 |

[a] A, original difference spectra; B, original difference spectra normalized to the ordinate at 1700 cm$^{-1}$; C: original difference spectra normalized to the area under the amide I band (1700–1600 cm$^{-1}$); D, original difference spectra normalized to both ordinate and area; E, same as D with the omission of the spectrum of trypsin inhibitor. [b] No. of factors corresponds to the total number of factors used for the regressions after elimination of those accounting for noise.

Table II: Effect of Spectral Range on Correlation Coefficients ($r$) for Helix, Sheet, and Turn Structures for the 18 Proteins

| spectral range (cm$^{-1}$) | $r$(helix) | $r$(sheet) | $r$(turns) | no. of factors |
|---|---|---|---|---|
| 1700–1600 | 0.995 | 0.989 | 0.903 | 11 |
| 1800–1500 | 0.975 | 0.976 | 0.819 | 8 |
| 1800–1600 | 0.971 | 0.984 | 0.656 | 9 |
| 1700–1500 | 0.978 | 0.933 | 0.857 | 8 |

Table III: Effect of Using Second-Derivative and Deconvolved Spectra on Correlation Coefficients ($r$) for Helix, Sheet, and Turn Structures for the 18 Proteins[a]

| | $r$(helix) | $r$(sheet) | $r$(turns) | no. of factors |
|---|---|---|---|---|
| D | 0.995 | 0.989 | 0.903 | 11 |
| D second derivative | 0.991 | 0.963 | 0.867 | 8 |
| D deconvolved | 0.978 | 0.944 | 0.687 | 7 |

[a] D is as defined in Table I. Second-derivative and deconvolved spectra were generated from normalized difference spectra as described in the text.

water or buffer spectra as appropriate, were used to create calibration sets for the subsequent CIRCOM analysis of unknowns. Following factor analysis, multiple linear regression establishes correlations between significant factors (i.e., those that do not account for noise) and the secondary structure of each standard. The correlation coefficients, together with the deviation sums of squares and an F-test when appropriate, may be used to identify the best calibration set for structural prediction. What follows is our identification of the best "type" of spectrum to present to CIRCOM in order to obtain the most reliable prediction of secondary structure.

When working with spacers of 6 $\mu$m, it is very difficult to obtain a reproducible path length because the cell has to be disassembled for cleaning. It seemed important, therefore, to make adjustments for any variations in baseline and absorbance due to variations in path length and concentration. We set the area under the amide I band (i.e., from 1700 to 1600 cm$^{-1}$) to be a constant and also set the ordinate at 1700 cm$^{-1}$ to a constant value by appropriate multiplication and offsetting of the spectra. The actual constants are not important, but the values obtained for papain were selected. It should be noted that we are assuming that the contribution of aromatic side chains to the absorption in this region is very small compared to the amide I absorption. Table I presents the correlation coefficients for $\alpha$-helical, $\beta$-sheet, and turn structures that were obtained after each of these normalizations for the 18 proteins. The number of factors represents the total number of factors used in the three regressions. The data show that normalization for the amide I band area increases the correlation coefficients for $\alpha$-helix and $\beta$-sheet (set C), and a further increase is obtained by the correction to a constant ordinate value at 1700 cm$^{-1}$ (set D). By contrast, the correlation coefficient for turns is decreased. However, by examining the deviation sums of squares and applying F-tests where appropriate (data not shown), it was determined that data set D was the most parsimonious model for each structure; that is, set D gave the most reliable prediction of each structure with the minimal number of terms in each regression (see below). Hence, set D was used as the basis for evaluating further spectral manipulations.

Another important parameter for consideration is the spectral range. Most IR studies of protein structure have concentrated on the amide I and II bands, and the most reliable structural assignments have been made for these bands as opposed to the other conformation-sensitive bands such as amide III and V. We investigated the effects of selecting various segments of the amide I/II region (1800–1500 cm$^{-1}$) on the correlation coefficients for the calibration set of 18 proteins. Spectra were normalized over the 1700–1600-cm$^{-1}$ region as above. The data, presented in Table II, show that the amide I region (1700–1600 cm$^{-1}$) gives the best correlation for each structure. This was confirmed by comparison of the deviation sum of squares and by applying F-tests.

Many of the literature assignments of the amide I and II bands of proteins have been made with the judicious use of second-derivative and deconvolution techniques for observing overlapping bands. We applied these algorithms to our standard spectra (after normalization) and generated calibration sets via CIRCOM by using the 1700–1600-cm$^{-1}$ region. The data shown in Table III indicate that the difference spectra give the best correlations for each structure. Again, this was confirmed by comparing deviation sums of squares and applying F-tests. We were careful not to overdeconvolve the spectra; the factors we used are lower than those typically used in qualitative analysis of these amide I bands.

Selection of the most appropriate X-ray analysis is also an important consideration in evaluating our correlations. Leaving aside the question of the relationship between crystal and solution structures, there is some debate among crystallographers on the definition of secondary structural elements from the X-ray data. The most comprehensive treatments of the X-ray data are those of Levitt and Greer (1977) and Kabsch and Sander (1983). Of the proteins used in our study, 16 are common to these original reports (ribonuclease A and chymotrypsinogen are absent from the latter). We therefore sought correlations between the IR data as analyzed by CIRCOM and each of these data sets. The correlations coefficients show that while the analysis for $\alpha$-helical structures is little different $r = 0.985$, Levitt and Greer, $n = 16$; $r = 0.983$, Kabsch and Sander, $n = 16$), a higher correlation for $\beta$-sheet is obtained by using the Levitt and Greer data ($r = 0.982$, $n = 16$) compared to the Kabsch and Sander data ($r = 0.927$, $n = 16$). These regressions could not be tested by the procedure used previously, since they are constructed from different data.

As a basis for analyzing the secondary structure of an unknown protein, our investigation shows that data set D (Table I) is the most appropriate presentation of our data to CIRCOM for generation of a calibration set. This data set will be used in the evaluation that follows. For set D, 11 of the original 18 factors remained after elimination of those accounting for noise. Of these, the loadings of six factors correlated with the composition of $\alpha$-helices in the standards. Similarly, eight factors correlated with $\beta$-sheet and three factors correlated with turns. Each factor may correlate with more than one property.

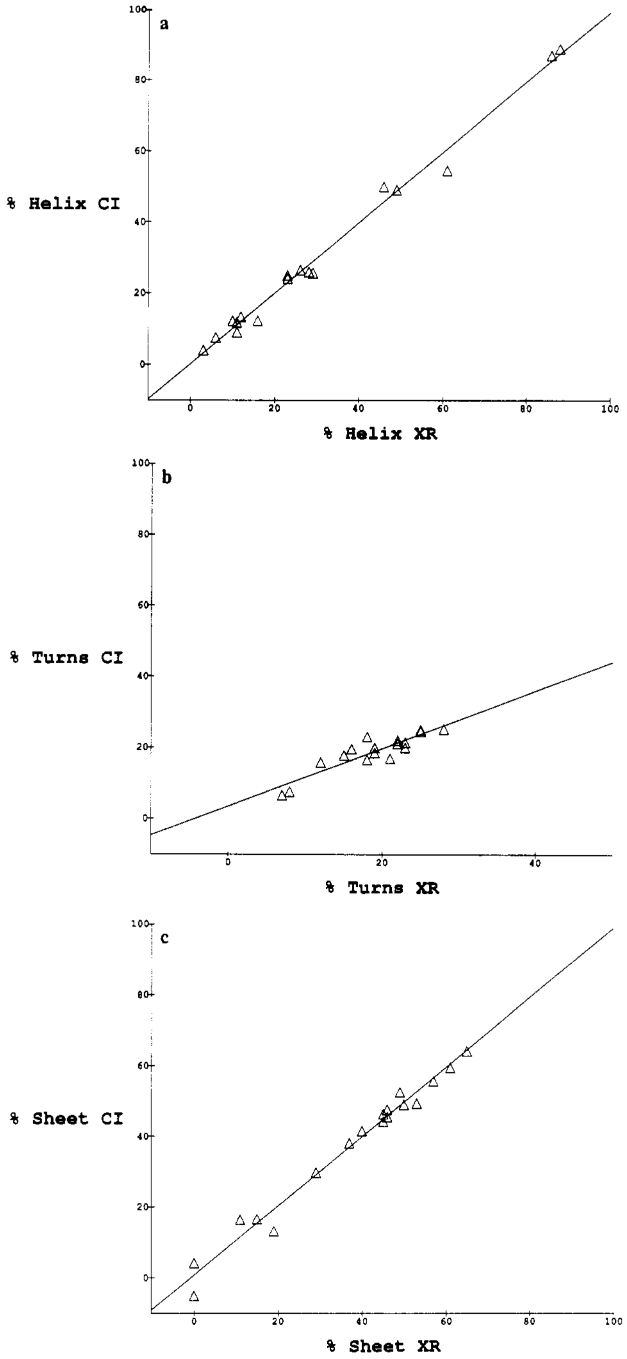

FIGURE 1: Correlation of secondary structure determined by CIRCOM with values from X-ray crystallography determined by Levitt and Greer (1977). X-ray values for each structure were entered into CIRCOM as true values; CIRCOM values were obtained from the calibration set containing all 18 standards. In (a), two points coincide exactly at $x = 26.0$, $y = 26.5$, and in (c), two points coincide exactly at $x = 23$, $y = 21.5$. (a) % $\alpha$-helix; (b) % $\beta$-sheet; (c) % turns. CI = CIRCOM; XR = X-ray.

Figure 1 presents the correlations for calibration set D in graphical form. In addition to showing the correlation, which is particularly good for $\alpha$-helix, these plots also serve to illustrate the range of secondary structure present in the calibration set.

*Prediction of Secondary Structure.* The correlations presented above have been obtained by multiple linear regression to ensure the best fit between the X-ray and CIRCOM values. In evaluating the usefulness of the method, it is important to check whether the predictions are accurate for samples outside the calibration set. Our choice of standards for the calibration set was influenced by the need to reflect the widest possible variation in secondary structural content. Nevertheless, the ideal method for validation would be to run the FTIR spectra of a prediction set equal in number to the calibration set and to examine the accuracy of the CIRCOM values. Given our limit of 18 spectra of known protein structures [according to Levitt and Greer (1977)], this would necessitate an unacceptably small calibration set of nine standards. We therefore chose to adopt an alternative method for validation whereby each protein, in turn, was eliminated from the analysis and a calibration set generated from the remainder was used to predict its secondary structure. The predictions of secondary structural content, via CIRCOM, are presented in Table IV together with the X-ray values (Levitt & Greer, 1977).

With our limited calibration set, it is important to test whether an unknown lies within the range of variation of properties convered by the standards. CIRCOM provides such a test via Mahalonobis' statistic ($D^2$) (Fredericks et al., 1985a,b). The average value for $D^2$ in the calibration set is $n_f/n_c$, where $n_f$ is the number of factors remaining after elimination of those accounting for noise and $n_c$ is the number of spectra. When all the factors are retained in the analysis, $D^2$ is 1 for every sample in the calibration set. As factors are eliminated, $D^2$ is reduced for each sample, with the average value for $D^2$ being $n_f/n_c$. For an unknown, if $D^2$ is significantly greater than $2n_f/n_c$ or 1 (whichever is the lesser), then this unknown can be judged to be outside the range of the calibration set and the prediction requires extrapolation of the regression equation(s). Values for $D^2$ are presented in Table IV. For our calibration set (set D), $2n_f/n_c = 1.22$ and we therefore choose $D^2 = 1$ as our cutoff point.

Examination of Table IV shows that in most cases agreement between the CIRCOM and X-ray values for our three classifications of secondary structure is good. Agreement is good in most of those cases for which $D^2$ is less than 1. Most of those cases in which there is a poorer match between the two values are identified by a large Mahalonobis statistic (i.e., myoglobin, hemoglobin, concanavalin A, insulin, protease, and prealbumin). It is of significance that these proteins are either mostly $\alpha$-helical or mostly $\beta$-sheet. Thus, when each one is removed from the calibration set, the range over which predictions are accurate is markedly reduced (especially for analyzing proteins with high helix or high sheet content). Using the full calibration set (18 rather than 17 standards), we would expect improved accuracy at these extremes. A Mahalonobis statistic of 1 cannot be considered an absolute cutoff since very accurate CIRCOM estimations are obtained for two proteins (papain and chymotrypsin) where $D^2$ is greater than 1. Additionally, the least accurate prediction (that for the trypsin inhibitor) gives a value for $D^2$ that is only slightly greater than 1. Nevertheless, when analyzing a complete unknown, the Mahalonobis statistic gives a guide to the identification of outliers.

The inaccuracy of the prediction for trypsin inhibitor almost certainly occurs as a result of its unusual FTIR spectrum. Examination of the second-derivative FTIR spectrum (our unpublished results) reveals strong bands at 1643 and 1661 cm$^{-1}$, which may be assigned to $\beta$-sheet and $\alpha$-helix/turn structures, respectively. The unusually high wavenumber of the $\beta$-sheet absorption probably arises from the strong twist of this structure in the protein. The effect in the difference spectrum is to produce an amide I maximum at 1657 cm$^{-1}$ with a strong shoulder at 1647 cm$^{-1}$. This is interpreted by CIRCOM as a higher proportion of $\alpha$-helical than $\beta$-sheet structure (almost an exact reversal of the X-ray values). Hence, it is

Table IV: Prediction of Protein Secondary Structure Using CIRCOM (Calibration Set D)[a]

| protein | $D^2$[b] | helix CI[c] | helix XR[d] | sheet CI | sheet XR | turns CI | turns XR |
|---|---|---|---|---|---|---|---|
| myoglobin | 1.61 | 94 | 88 | −11 | 0 | 4 | 7 |
| hemoglobin | 1.27 | 76 | 86 | 16 | 0 | 4 | 8 |
| insulin | 1.35 | 53 | 61 | 20 | 15 | 14 | 12 |
| cytochrome *c* | 0.72 | 53 | 49 | 17 | 11 | 17 | 22 |
| lysozyme | 0.89 | 54 | 46 | 7 | 19 | 19 | 23 |
| alcohol dehydrogenase | 1.01 | 26 | 29 | 40 | 40 | 19 | 19 |
| papain | 1.72 | 23 | 28 | 34 | 29 | 22 | 18 |
| trypsin inhibitor | 1.06 | 46 | 26 | 26 | 45 | 13 | 16 |
| nuclease | 0.82 | 32 | 26 | 36 | 37 | 22 | 23 |
| ribonuclease A | 0.72 | 19 | 23 | 49 | 46 | 16 | 21 |
| ribonuclease S | 0.72 | 18 | 23 | 49 | 53 | 21 | 15 |
| carbonic anhydrase | 0.76 | 6 | 16 | 54 | 45 | 23 | 25 |
| chymotrypsinogen | 0.91 | 20 | 12 | 45 | 49 | 24 | 23 |
| chymotrypsin | 1.20 | 13 | 11 | 50 | 50 | 22 | 25 |
| protease | 1.35 | 13 | 11 | 49 | 57 | 24 | 18 |
| elastase | 0.97 | 11 | 10 | 51 | 46 | 21 | 28 |
| prealbumin | 1.69 | 11 | 6 | 40 | 61 | 25 | 19 |
| concanavalin A | 1.61 | 13 | 3 | 57 | 65 | 28 | 22 |

[a] The accuracy of the CIRCOM method was validated by removing each protein, in turn, from the calibration set (set D) and using the remaining 17 spectra to predict its composition. [b] $D^2$ = Mahalonobis' statistic (see text). [c] CI = CIRCOM estimations. [d] XR = X-ray values (Levitt & Greer, 1977).

Table V: Prediction of Protein Secondary Structure Using CIRCOM (Calibration Set E)[a]

| protein | $D^2$[b] | helix CI[c] | helix XR[d] | sheet CI | sheet XR | turns CI | turns XR |
|---|---|---|---|---|---|---|---|
| myoglobin | 1.58 | 86 | 88 | −8 | 0 | 15 | 7 |
| hemoglobin | 1.37 | 81 | 86 | 16 | 0 | 12 | 8 |
| insulin | 1.16 | 53 | 61 | 21 | 15 | 23 | 12 |
| cytochrome *c* | 0.72 | 53 | 49 | 21 | 11 | 15 | 22 |
| lysozyme | 0.89 | 55 | 46 | 5 | 19 | 13 | 23 |
| alcohol dehydrogenase | 0.94 | 30 | 29 | 44 | 40 | 20 | 19 |
| papain | 1.59 | 26 | 28 | 32 | 29 | 21 | 18 |
| nuclease | 0.70 | 26 | 26 | 36 | 37 | 26 | 23 |
| ribonuclease A | 0.87 | 25 | 23 | 38 | 46 | 16 | 21 |
| ribonuclease S | 0.77 | 23 | 23 | 38 | 53 | 24 | 15 |
| carbonic anhydrase | 0.76 | 12 | 16 | 52 | 45 | 24 | 25 |
| chymotrypsinogen | 0.91 | 16 | 12 | 46 | 49 | 21 | 23 |
| protease | 1.06 | 15 | 11 | 50 | 57 | 16 | 18 |
| chymotrypsin | 0.69 | 11 | 11 | 50 | 50 | 22 | 25 |
| elastase | 0.84 | 9 | 10 | 55 | 46 | 18 | 28 |
| prealbumin | 1.38 | 2 | 6 | 61 | 61 | 28 | 19 |
| concanavalin A | 1.48 | 2 | 3 | 70 | 65 | 30 | 22 |

[a] The accuracy of the CIRCOM method was validated by removing each protein, in turn, from the calibration set (set E) and using the remaining 16 spectra to predict its composition. [c] $D^2$ = Mahalonobis' statistic (see text). [c] CI = CIRCOM estimations. XR = X-ray values (Levitt & Greer, 1977).

possible that the inclusion of trypsin inhibitor, with its unusual spectrum, in the calibration set may lead to inaccuracies in the predictions of more typical structures. We therefore generated another calibration set of 17 proteins (set E, omitting trypsin inhibitor from set D) and performed the validation procedure as outlined above. The correlation data are given in Table I, while the results of the validation procedure are presented in Table V. For data set E, nine of the original 17 factors remained after discarding those accounting for noise. Of these, the loadings of three factors correlated with the $\alpha$-helical composition of the standards. Similarly, two factors correlated with $\beta$-sheet and eight factors with turns. Again, each factor may correlate with more than one property.

Adoption of this calibration set resulted in an improvement in the accuracy of the predictions for $\alpha$-helical structure (Table V). In order to compare the accuracy of different methods for predicting each type of secondary structure, we calculate the standard error of prediction (SEP):

$$\mathrm{SEP} = \left[ \frac{\sum_{j}^{n} (p_{c_j} - p_{x_j})^2}{n} \right]^{1/2}$$

where $p_{c_j}$ = the proportion of structure predicted by CIRCOM in protein *j*, $p_{x_j}$ = the proportion of structure calculated by Levitt and Greer (1977) from the original X-ray data for protein *j*, and *n* = the number of proteins. The SEP (in % structure) for $\alpha$-helix falls from 7.8 (set D) to 3.9 (set E), the SEP for $\beta$-sheet decreases from 9.7 (set D) to 8.3 (set E), and the SEP for turns increases from 4.3 (set D) to 6.6 (set E).

CIRCOM also provides a comparison of the original difference spectrum with that generated as the products of the required factors and factor loadings necessary to produce the best fit. This allows a visual comparison of the goodness of fit and therefore the accuracy of the prediction. Examples of these spectra are shown in Figure 2. In most cases, where $D^2$ is less than 1, the fit is extremely good. In the case of hemoglobin, the fit is at least as good as that previously obtained by curve-fitting of amide I components revealed by deconvolution (Byler & Susi, 1986).

CIRCOM also provides eigenspectra for each structural type. Each spectrum is generated by using the factors retained in the regression equation, the corresponding factor loadings for each standard, and the regression coefficient for the calibration set. Thus, each eigenspectrum represents the sum of factors in the correct proportion to generate the spectrum of 100% $\alpha$-helix, $\beta$-sheet, or turn as defined by the spectra in the calibration set. This approach may be used to determine which regions of the spectrum correlate with the presence of each structure. The eigenspectra (not shown) constructed from set E show "bands" at 1695, 1659 (major), and 1641 $\mathrm{cm}^{-1}$ for $\alpha$-helix, bands at 1682 and 1634 $\mathrm{cm}^{-1}$ (major) for $\beta$-sheet, and

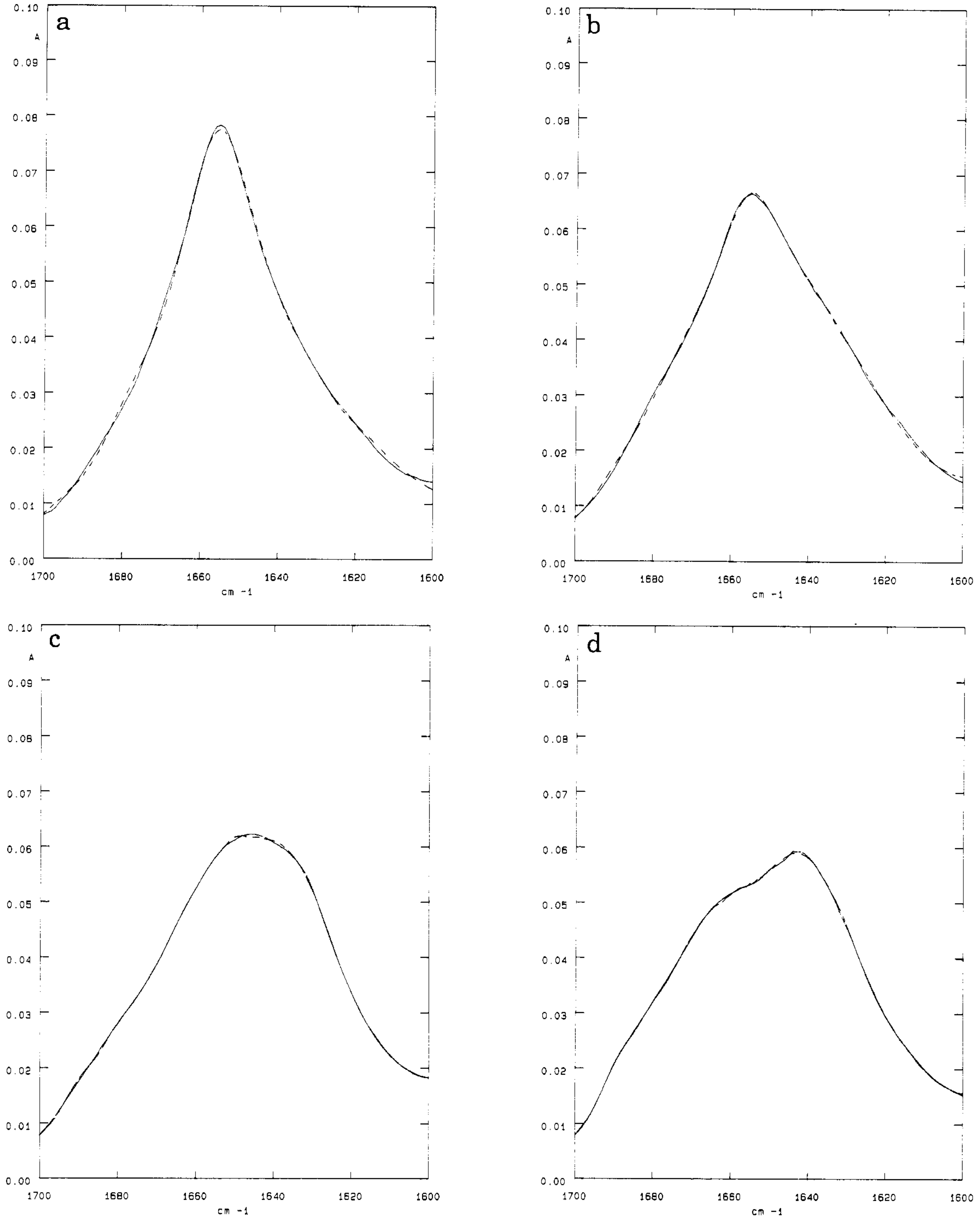

FIGURE 2: Normalized difference spectra of selected proteins in $H_2O$ (solid lines) compared with spectra reconstructed from CIRCOM factors and factor loadings to produce the best fit (broken lines): (a) hemoglobin; (b) cytochrome *c*; (c) alcohol dehydrogenase; (d) ribonuclease A.

Table VI: Prediction of Secondary Structure for Proteins Not Analyzed by Levitt and Greer (1977)[a]

| protein | $D^2$[b] | helix CI[c] | helix XR[d] | sheet CI | sheet XR | turns CI | turns XR |
|---|---|---|---|---|---|---|---|
| calmodulin | 1.01 | 67 | 63 | 10 | 11 | 5 | |
| pepsin | 1.07 | 24 | 12 | 43 | 46 | 20 | 22 |
| trypsin | 0.57 | 14 | 8 | 47 | 59 | 25 | 23 |
| trypsinogen | 0.62 | 18 | | 45 | | 23 | |
| phospholipase $A_2$ | 0.94 | 47 | 48 | 25 | 17 | 23 | 20 |

[a] Calibration set E was used. [b] $D^2$ = Mahalonobis' statistic. [c] CI = CIRCOM estimations. [d] XR = X-ray values (see text).

bands at 1684 (major), 1654, and 1633 cm$^{-1}$ for turns.

A series of proteins for which X-ray data are available but which were not analyzed by Levitt and Greer (1977) were also investigated by CIRCOM. The results are presented in Table VI together with estimations of secondary structure made by the original X-ray workers (where available). The CIRCOM estimations of secondary structural content shown were made with set E as the calibration set. For the proteins in Table VI, the use of set E gave a lower Mahalonobis statistic than set D, which implies that these are the more accurate predictions. Additionally, the secondary structure predictions made with set D (data not shown) were much less accurate

Table VII: Prediction of Secondary Structure of Proteins in $D_2O$ Using CIRCOM[a]

| protein | $D^2$[b] | helix CI[c] | helix XR[d] | sheet CI | sheet XR | turns CI | turns XR |
|---|---|---|---|---|---|---|---|
| myoglobin | 2.46 | 128 | 88 | 7 | 0 | 14 | 7 |
| hemoglobin | 0.75 | 63 | 86 | 9 | 0 | 18 | 8 |
| cytochrome *c* | 0.94 | 51 | 49 | 12 | 11 | 20 | 22 |
| lysozyme | 1.00 | 59 | 46 | 5 | 19 | 16 | 23 |
| papain | 2.34 | 10 | 28 | 23 | 29 | 34 | 18 |
| ribonuclease A | 0.67 | 13 | 23 | 48 | 46 | 22 | 21 |
| ribonuclease S | 0.76 | 16 | 23 | 49 | 53 | 22 | 15 |
| carbonic anhydrase | 0.58 | 16 | 16 | 53 | 45 | 21 | 25 |
| chymotrypsinogen | 0.78 | 24 | 12 | 45 | 49 | 25 | 23 |
| chymotrypsin | 2.49 | 10 | 11 | 50 | 50 | 36 | 25 |

[a] The accuracy of the CIRCOM method was validated by removing each protein, in turn, from the calibration set and using the remaining nine spectra to predict its composition. [b] $D^2$ = Mahalonobis' statistic. [c] CI = CIRCOM estimations. [d] XR = X-ray values (Levitt & Greer, 1977).

(SEP = 9.1, $n$ = 11) than those made via calibration set E (SEP = 6.3, $n$ = 11). These SEP values are composites of the predictions for the three structures. Estimations of secondary structure from X-ray data were made from the original reports as follows: calmodulin (Babu et al., 1985), pepsin (Andreeva et al., 1984), trypsin-DIP (Stroud et al., 1972; Levitt & Greer, 1977), and phospholipase $A_2$ (Dijkstra et al., 1981, 1983). The estimations obtained via CIRCOM are in reasonable agreement with the values given by the original X-ray workers.

### *(B) Spectra in $^2H_2O$*

*Generation of Calibration Set.* A calibration set consisting of FTIR spectra of proteins in $^2H_2O$ was generated by using the same criteria as those discussed above. Again, the area under the amide I band (1700–1600 $cm^{-1}$) and the absorbance value at 1700 $cm^{-1}$ were normalized in all spectra. However, only 10 spectra of proteins in $^2H_2O$ could be used for a calibration set. This was because of difficulties in reaching complete exchange for several of the proteins studied. In some cases, elevated temperatures were used to increase the extent of exchange, but control spectra of the same proteins in $H_2O$ revealed band shifts indicating conformational changes after heating. In other cases, extensive incubation at room temperature resulted in protein precipitation and denaturation as revealed by the FTIR spectrum. Proteins for which we encountered the above problems were eliminated from the calibration set.

Correlation coefficients for the proteins selected for the calibration set were 0.964 for $\alpha$-helix, 0.970 for $\beta$-sheet, and 0.944 for turns. For comparison, a calibration set using spectra of the same proteins in $H_2O$ gave correlation coefficients of 0.998 ($\alpha$-helix), 0.990 ($\beta$-sheet), and 0.943 (turns).

*Prediction of Secondary Structure.* In order to assess the accuracy of this calibration set for predicting the secondary structure of "unknown" proteins, the method was validated by using the same procedure as used for spectra of proteins in $H_2O$ (see Table IV). The results of this validation are presented in Table VII. This validation was compared with a validation carried out on a calibration set comprising the spectra of the same 10 proteins in $H_2O$. For $\alpha$-helix, SEP($H_2O$) = 8.5 and SEP($^2H_2O$) = 17.1, for $\beta$-sheet, SEP($H_2O$) = 7.4 and SEP($^2H_2O$) = 6.8, and for turns, SEP($H_2O$) = 5.8 and SEP($^2H_2O$) = 8.1. Therefore, while the accuracy for $\beta$-sheet is little different, spectra of samples in $H_2O$ give more accurate predictions for $\alpha$-helix and turns.

## Discussion

We have developed an accurate, reliable, and reproducible infrared method for determining the secondary structure content (in terms of % $\alpha$-helix, % $\beta$-sheet, and % turns) of proteins in water. Validation of the appropriate calibration set (set E) has shown that we may expect a standard error of prediction of 3.9% for $\alpha$-helix, 8.3% for $\beta$-sheet, and 6.6% for turns. Assessment of the method using proteins that have not been characterized by Levitt and Greer (1977) confirms these accuracies. The SEP values compare well with those obtained from other spectroscopic methods used for quantitative estimations of protein secondary structure. Comparable SEP values from circular dichroism for $\alpha$-helix, $\beta$-sheet, and turn are 5%, 6%, and 10% (Provencher & Glockner, 1981) and 6%, 7%/5% (antiparallel/parallel), and 5% (Hennessey & Johnson, 1981), respectively. A recent method using Raman spectroscopy (Bussian & Sander, 1989) obtains accuracies of 6% for $\alpha$-helix and 5% for $\beta$-sheet, which are comparable with the earlier Raman methods of Williams (1983, 1986). The infrared method of Byler and Susi (1986) produces SEP values of 2.2% and 2.7% for $\alpha$-helix and $\beta$-sheet, respectively.

The CIRCOM method that has been developed from the published procedures of Fredericks et al. (1985a,b) offers important advantages for the analysis of protein structure when compared to the commonly used method of deconvolution followed by curve-fitting (Byler & Susi, 1986; Surewicz & Mantsch, 1988). First, no deconvolution (or generation of derivatives) is required, and pretreatment of the data is kept to a minimum. It is generally accepted that deconvolution procedures should be handled with care. Overdeconvolution produces negative side lobes on absorption bands and can result in artificial bands from noise or incomplete compensation of water vapor (Mantsch et al., 1988). Thus, several distortions may be induced in the spectrum before curve-fitting is applied. There is a further problem concerning the consistent application of this technique between laboratories.

A second advantage of the CIRCOM method is that no assignment of the amide I components is necessary. This is an important feature, since there are several instances where an assignment of all the bands is impossible to make with complete certainty. In particular, it is difficult to differentiate the high-wavenumber component associated with antiparallel $\beta$-sheet from several absorptions associated with $\beta$-turns (Haris et al., 1986; Surewicz & Mantsch, 1988). In cases where there is some doubt in assignment, the spectroscopist may be tempted to sum the curve-fitted components according to preconceptions of the "correct" secondary structure. Further difficulties arise with the assignment to amide I components of bands due to noise or water vapor (above), amino acid side chains, or contaminants. The CIRCOM method should eliminate the influence of the first two, since they should be present in the standards and will not correlate with secondary structure. Those factors that account for noise and any water vapor will be more readily identified by CIRCOM when they are significantly smaller than the factors accounting for structural detail. Obviously, all methods rely for accuracy on the highest obtainable purity for standards and unknowns.

A third difficulty with the deconvolution/curve-fitting approach lies in inaccuracies in the curve-fitting step. Typically, only those components observed by deconvolution are used to regenerate the band shape of either the original or deconvolved (Byler & Susi, 1986) spectrum. In our experience, this results in an incomplete fit and a nonunique solution, casting doubt on the accuracy of the quantitative determination derived from the band areas of the components. A better fit can usually be obtained by adding weaker components not revealed by deconvolution. While it is certain that neither derivative nor deconvolved spectra reveal all the component bands, the user would have little justification in accepting and assigning these extra bands.

An assumption common to earlier methods and our method is that the extinction coefficients of the bands assigned to $\alpha$-helical, $\beta$-sheet, and turn structures are identical. There seems little reason to accept this, rather the opposite; evidence for changes in extinction coefficients as proteins undergo conformational changes is now emerging (Jackson et al., 1989; Mantsch et al., 1989). This is also a weakness in our approach, since we have normalized the area under the amide I band in order to account for variations in path length and concentration. However, since the normalization produces better correlations, it appears that variations in path length are more important than variations in absorption coefficient. We are also assuming that the contribution of amino acid side-chain absorption to the 1700–1600-cm$^{-1}$ region is relatively small. While this is known to be the case (Chirgadze et al., 1975), the contribution will vary according to primary structure. However, we would not expect the factors accounting for these contributions to correlate with secondary structure and they should therefore be eliminated at the regression step. This assumption is common to other methods, since the areas of the band assigned to different secondary structures are expressed as a percentage of the total area under the amide I band.

When a method is used that provides a direct report of structural content with little user intervention, it is obviously important to minimize the likely sources of error. In particular, many infrared spectroscopists have avoided the use of water as a solvent because of difficulties in the subtraction of the H–O–H deformation band, which overlaps with the amide I band and is much more intense at the protein concentrations of interest to the biochemist. In our experience, a digital subtraction of the water band can be successfully judged by eye, by using the criteria of a level baseline from 1900–1720 cm$^{-1}$ and the elimination of the water band near 2130 cm$^{-1}$. Nevertheless, recent authors have investigated the use of algorithms for obtaining the best subtraction factor for difference spectroscopy (Powell et al., 1986; Dousseau et al., 1989). These methods obtain very good reproducibility between repeat spectra, but it is of significance that Dousseau et al. find it impossible to counter variations in path length induced by cleaning of 6-$\mu$m cells. As a result, they normalize the area under the amide I band in order to obtain better ordinate reproducibility. This procedure is identical with that found to be necessary by us in optimizing our correlation coefficients. In principle, any variation in the accuracy of the subtraction of the water band will be countered by the use of factor analysis and multiple linear regression to identify those eigenspectra that correlate with variations in secondary structure of the calibration set. The accuracy of the subtraction is obviously independent of protein structure, and we can expect the factor(s) that account for this to be eliminated in the regression step.

Our careful investigation of the criteria for the construction of an accurate calibration set raises a number of important points. First, the identification of the amide I region alone as being the most reliable indicator for quantitative assessments of secondary structure is certainly a reflection of earlier qualitative studies that have successfully made structural assignments for this region. The loss of accuracy found when including the amide II band in the calibration probably reflects spectral contributions from amino acid side chains in this region (Chirgadze et al., 1975). Second, the reduction in correlation coefficients when second-derivative and, more particularly, deconvolved spectra were used is possibly a consequence of the corruption and loss of significant spectral information by these algorithms. We were conservative in our selection of deconvolution factors, as it is known that over-deconvolution produces distortions in baseline and band shapes. This suggests that quantitative interpretation of curve-fitting to deconvolved spectra (Byler & Susi, 1986) should be made with considerable caution [as noted by Surewicz and Mantsch (1988)]. Third, our finding that the IR spectra correlate better with the interpretation of secondary structure from the X-ray data as developed by Levitt and Greer (1977) than that of Kabsch and Sander (1983) is in agreement with other workers (Byler & Susi, 1986). The method of Kabsch and Sander is essentially based on the recognition of hydrogen-bonding patterns, whereas that of Levitt and Greer is based on $C_\alpha$ coordinates. As the position of the amide I band is known to be closely dependent on hydrogen-bonding, we might expect better correlation with the former. However, a higher proportion of primary sequence is assigned to secondary structure by the method of Levitt and Greer, and this probably explains the closer correlation with the spectroscopic data.

The use of Mahalonobis' statistic is an important feature of the method. The statistic, generated for each prediction, gives an indication for outliers, i.e., those spectra for which the variation in selected properties lies outside the range defined in the calibration set. In these cases the spectrum should be reanalyzed with an alternative calibration set. This has important implications for the analysis of, for example, polypeptides and membrane proteins, where the variation in spectroscopic properties with secondary structure can be expected to be different than that described by a calibration set constructed from water-soluble globular proteins.

The removal of the spectrum of bovine pancreatic trypsin inhibitor in the calibration set demonstrates another important feature, since the accuracy of the predictions of $\alpha$-helix and $\beta$-sheet are improved but that for turns is reduced. Thus, future development of this method could proceed via the generation of separate calibration sets for these properties. An initial prediction could be made by using a wide-ranging set, and this prediction could be used to select a second calibration set whose properties varied more narrowly over the region of interest and would be expected to give a more accurate prediction. The use of this approach obviously requires more spectral standards than used by us in this initial report.

Our method is less successful in analyzing spectra of proteins in $^2H_2O$. Although in these cases there is no longer any variability induced by the subtraction factor for solvent, there is a greater problem caused by variations in the extent of hydrogen–deuterium exchange. Our observation of no further spectroscopic changes in the amide II region ensures only that exchange has stabilized, but in many cases it may not be complete. As the spectra of many proteins contain amino acid side-chain absorptions that overlap with the amide II band (Chirgadze et al., 1975), it is difficult to determine the extent of exchange from the intensity of the amide II band. A more reliable approach would be to monitor the disappearance of the amide A band near 3300 cm$^{-1}$ (Mitchell et al., 1988). The practical difficulties in ensuring complete or equivalent exchange states also makes the application of quantitative methods based on curve-fitting very uncertain, since several band shifts in the amide I region are known to occur on exchange.

Although the predictions generated by the CIRCOM method are very good, there remains scope for improvement. We have already noted that a larger number of standards of known structure in the calibration set will improve the accuracy of the prediction of an unknown. A single large calibration set or sets specific to narrow ranges of secondary structure could be used. Obviously, the greater the number of standards, the greater the likelihood of accounting for all possible variations in structural type. We stress that, in the analysis of a complete unknown, this method should be used in conjunction with a qualitative analysis using derivative and/or deconvolution methods, since these spectra will alert the spectroscopist to possible discrepancies. If the positions of the bands assigned to the various secondary structures are within their expected ranges, then the quantitative assessment obtained via factor analysis may be viewed with confidence.

## Acknowledgments

We thank Dorian Kennedy (Royal Free Hospital School of Medicine) for obtaining the spectrum of phospholipase $A_2$ and Brian Bond (SmithKline Beecham Pharmaceuticals) for assistance with statistical analysis.

## References

Andreeva, N. S., Zdanov, A. S., Gustchina, A. E., & Fedorov, A. A. (1984) *J. Biol. Chem. 259*, 11353–11365.

Arrondo, J. L. R., Young, N. M., & Mantsch, H. H. (1988) *Biochim. Biophys. Acta 952*, 261–268.

Babu, S. Y., Sack, J. S., Greenhough, T. S., Bugg, C. E., Means, A. R., & Cook, W. J. (1985) *Nature 315*, 37–40.

Bussian, B. M., & Sander, C. (1989) *Biochemistry 28*, 4271–4277.

Byler, D. M., & Susi, H. (1986) *Biopolymers 25*, 469–487.

Chirgadze, Yu. N., Fedorov, O. V., & Trushina, N. P. (1975) *Biopolymers 14*, 679–694.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J., & Drenth, J. (1981) *J. Mol. Biol. 147*, 97–123.

Dijkstra, B. W., Renetseder, R., Kalk, K. H., Hol, W. G. J., & Drenth, J. (1983) *J. Mol. Biol. 168*, 163–179.

Dousseau, F., Therrien, M., & Pezolet, M. (1989) *Appl. Spectrosc. 40*, 538–542.

Eckert, M., Grosse, R., Malur, J., & Repke, K. R. M. (1977) *Biopolymers 16*, 2549–2563.

Fredericks, P. M., Lee, J. R., Osborn, P. R., & Swinkels, D. A. J. (1985a) *Appl. Spectrosc. 39*, 303–310.

Fredericks, P. M., Lee, J. R., Osborn, P. R., & Swinkels, D. A. J. (1985b) *Appl. Spectrosc. 39*, 311–316.

Haris, P. I., Lee, D. C., & Chapman, D. (1986) *Biochim. Biophys. Acta 874*, 255–265.

Hennessey, J. P., & Johnson, W. C. (1981) *Biochemistry 20*, 1085–1094.

Holloway, P. W., & Mantsch, H. H. (1989) *Biochemistry 28*, 931–935.

Jackson, M., Haris, P. I., & Chapman, D. (1989) *Biochim. Biophys. Acta 985*, 75–79.

Kabsch, W., & Sander, S. (1983) *Biopolymers 22*, 2577–2637.

Kauppinen, J. K., Moffat, D. J., Mantsch, H. H., & Cameron, D. G. (1981) *Appl. Spectrosc. 35*, 271–276.

Koenig, J. L., & Tabb, D. L. (1980) in *Analytical Applications of FTIR to Molecular and Biological Systems* (Durig, J. R., Ed.) pp 241–255, Reidel, Holland.

Lee, D. C., & Chapman, D. (1986) *Biosci. Rep. 6*, 235–256.

Lee, D. C., Herzyk, E., & Chapman, D. (1987) *Biochemistry 26*, 5775–5783.

Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–293.

Maddams, W. F., & Southon, M. J. (1982) *Spectrochim. Acta 38A*, 459–466.

Malinowski, E. R., & Howery, D. G. (1980) *Factor Analysis in Chemistry*, Wiley, New York.

Mantsch, H. H., Moffat, D. J., & Casal, H. L. (1988) *J. Mol. Struct. 173*, 285–298.

Mantsch, H. H., Surewicz, W. K., Muga, A., Moffat, D. J., & Casal, H. L. (1989) 7th International Conference on Fourier Transform Spectroscopy, Fairfax, VA, 1989 (Cameron, D. G., Ed.) Proceedings of SPIE—The International Society for Optical Engineering, Vol. 1145, pp 580–581, SPIE, Bellingham, WA.

Mitchell, R. C., Haris, P. I., Fallowfield, C., Keeling, D. J., & Chapman, D. (1988) *Biochim. Biophys. Acta 941*, 31–38.

Powell, J. R., Wasacz, F. M., & Jakobsen, R. J. (1986) *Appl. Spectrosc. 40*, 339–344.

Provencher, S. W., & Glockner, J. (1981) *Biochemistry 20*, 33–37.

Ruegg, M., Metzger, V., & Susi, H. (1975) *Biopolymers 14*, 1465–1471.

Savitzky, A., & Golay, M. J. E. (1964) *Anal. Chem. 36*, 1627–1639.

Stroud, R. M., Kay, L. M., & Dickerson, R. E. (1972) *Cold Spring Harbor Symp. Quant. Biol. 36*, 125–140.

Surewicz, W. K., & Mantsch, H. H. (1988) *Biochim. Biophys. Acta 952*, 115–140.

Surewicz, W. K., Moscarello, M. A., & Mantsch, H. H. (1987) *J. Biol. Chem. 262*, 8598–8602.

Susi, H. (1969) in *Structure and Stability of Biological Molecules* (Timasheff, S. N., & Fasman, G. D., Eds.) pp 573–633, Marcel Dekker, New York.

Susi, H., & Byler, D. M. (1983) *Biochem. Biophys. Res. Commun. 115*, 391–397.

Susi, H., Byler, D. M., & Purcell, J. M. (1985) *J. Biochem. Biophys. Methods 11*, 235–240.

Susi, H., Timasheff, S. N., & Stevens, L. (1967) *J. Biol. Chem. 242*, 5460–5466.

Timasheff, S. N., Susi, H., & Stevens, L. (1967) *J. Biol. Chem. 242*, 5467–5473.

Williams, R. W. (1983) *J. Mol. Biol. 166*, 581–603.

Williams, R. W. (1986) *Methods Enzymol. 130*, 311–331.

Yang, P. W., Mantsch, H. H., Arrondo, J. L. R., Saint-Girons, I., Guillou, Y., Cohen, G. N., & Barzu, O. (1987) *Biochemistry 26*, 2706–2711.